WRF Precipitation Performance and Predictability for Systematically Varied Parameterizations over Complex Terrain

JULIA JEWORREK,^a GREGORY WEST,^b AND ROLAND STULL^a

^a *The University of British Columbia, Vancouver, British Columbia, Canada*

^b *BC Hydro, Vancouver, British Columbia, Canada*

(Manuscript received 21 October 2020, in final form 9 February 2021)

ABSTRACT: Physics parameterizations in the Weather Research and Forecasting (WRF) Model are systematically varied to investigate precipitation forecast performance over the complex terrain of southwest British Columbia (BC). Comparing a full year of modeling data from over 100 WRF configurations to station observations reveals sensitivities of precipitation intensity, season, location, grid resolution, and accumulation window. The choice of cumulus and microphysics parameterizations is most important. The WSM5 microphysics scheme yields competitive verification scores when compared to more sophisticated and computationally expensive parameterizations. Although the scale-aware Grell–Freitas cumulus parameterization performs better for summertime convective precipitation, the conventional Kain–Fritsch parameterization better simulates wintertime frontal precipitation, which contributes to the majority of the annual precipitation in southwest BC. Finer grid spacings have lower relative biases and a more realistic spread in precipitation intensity distribution, yet higher relative standard deviations of their errors—they produce finer spatial differences and local extrema. Finer resolutions produce the best fraction of correct-to-incorrect forecasts across all precipitation intensities, whereas the coarser 27-km domain yields the highest hit rates and equitable threat scores. Verification metrics improve greatly with longer accumulation windows—hourly precipitation values are prone to double-penalty issues, while longer accumulation windows compensate for timing errors but lose information about short-term precipitation intensities. This study provides insights regarding WRF precipitation performance in complex terrain across a wide variety of configurations, using metrics important to a range of end users.

SIGNIFICANCE STATEMENT: Numerical models are the most common tool to predict the weather. They rely on several physics packages, some of which approximate processes that are relatively small compared to the model grid size, while others provide the model with important ground surface properties. Since weather characteristics can be very different across Earth, the ideal combination of these physics packages varies by region. We sought the optimal model setup to predict precipitation in southwest British Columbia where rainfall is strongly influenced by seasons and mountains. This study demonstrates general limitations with precipitation forecasts. For example, short and extreme rainfall is generally difficult to predict. Forecasts in front of and over mountain ranges are often too wet, whereas forecasts behind mountain ranges are often too dry. We identified different model setups that performed best in the summer dry season and the cool wet season. Our results can inform forecasters regarding better model setups in areas with similar weather characteristics to yield better precipitation forecasts.

KEYWORDS: Precipitation; Forecast verification/skill; Model evaluation/performance; Numerical analysis/modeling; Parameterization; Complex terrain

1. Background and introduction

The climate of southwestern British Columbia (BC), Canada, is dominated by two hydrometeorological seasons: a wet, cool season (fall, winter, spring) with a persistent westerly storm track from the Pacific Ocean, and a drier, warm summer season with semipermanent high pressure off the coast (Mass 2008). The Vancouver Island Ranges and the Coast Range see enhanced precipitation on windward slopes due to orographic lifting, and a rain-shadow effect on the lee side. Abundant cool season precipitation provides the BC south coast with reliable water for drinking and generating hydroelectric power. However, lasting

and heavy rainfall in the cold season can also cause flooding and requires careful management of reservoirs. Therefore, accurate precipitation forecasts are crucial for resource management, risk assessment, and disaster mitigation.

Predictability with numerical weather prediction (NWP) is generally limited by imperfect initial conditions and simplified model approximations. In BC these forecasting challenges are further complicated: spatial observations from satellites or radar are not reliable or are partially blocked across complex terrain (Derin and Yilmaz 2014; Maggioni et al. 2016; Cookson-Hills et al. 2017; Sun et al. 2018), and station measurements sample the area unevenly due to the inaccessible terrain and the uneven population density and associated weather-station locations. The paucity of in situ observations upstream (i.e., the Pacific “data void”; Stull et al. 2004; Mass 2008) and across BC reduces the reliability of the initialization as well as the ability to evaluate fine differences between models comprehensively. Orographic clouds forced by the

Denotes content that is immediately available upon publication as open access.

Corresponding author: Julia Jeworrek, jjeworrek@eoas.ubc.ca

steep topography have complex mixed-phase microphysical processes, which are challenging to represent in NWP simulations (Colle et al. 2005; Rauber et al. 2019). Moreover, complex terrain can cause grid distortion and induce numerical error through artificial model behavior. Also, steep slopes redirect fast horizontal winds into fast vertical winds, which violate the Courant–Friedrichs–Lewy condition (i.e., CFL error; Courant et al. 1928) in the vertical. Another problem at common NWP model resolutions is the terrain smoothing applied for numerical stability, which misrepresents the true altitudes of mountain tops and valleys, and can cause false advection or blocking (Klemp 2011; Chow et al. 2019; Wiersema et al. 2020). Accurate representation of topography is important to correctly force the orographic enhancement of precipitation and potential rain shadow and foehn effects, which frequently determine the precipitation patterns in southwest BC.

Precipitation in NWP is a subgrid-scale process and a direct product of 1) a microphysics parameterization, which represents processes that control the formation, growth, and fallout of hydrometeors from clouds, and 2) a cumulus parameterization, which represents the effect of unresolved vertical motion on the grid variables. However, vertical mixing and turbulence fluxes estimated by a planetary boundary layer (i.e., turbulence) parameterization are also important to simulate the wind, stability, and moisture conditions that lead to precipitation (Pohl et al. 2011; Di Luca et al. 2014; Pei et al. 2014; Meynadier et al. 2015). The land surface model controls soil moisture and heat fluxes to provide the atmospheric model with heat, moisture, and radiation input from the ground, which affect the input variables for other parameterizations. Therefore, the choice of land surface model can also have a significant impact on the atmospheric water cycle and the diurnal precipitation cycle (Fan 2009; Duda et al. 2017; Wong et al. 2020) but is less often investigated in NWP sensitivity studies.

Some of these studies (e.g., Liu et al. 2011; Toride et al. 2019) show substantial sensitivity of precipitation to microphysics schemes; among the best schemes they found were Thompson (Liu et al. 2011; Rajeevan et al. 2010) and/or Morrison (Liu et al. 2011; Orr et al. 2017; Pu et al. 2019). However, other studies (Argüeso et al. 2011; Conrick and Mass 2019; Jeworrek et al. 2019; García-Díez et al. 2015; Sikder and Hossain 2016; Zeyaeyan et al. 2017; Hu et al. 2018) showed little sensitivity to microphysics schemes, with some investigators recommending Thompson (Conrick and Mass 2019) or WSM5 (Jeworrek et al. 2019).

In other studies, instead of microphysics, the cumulus parameterization was the most critical differentiator for precipitation skill, especially in convective development (e.g., Jankov et al. 2005; Pérez et al. 2014; Di Luca et al. 2014; García-Díez et al. 2015; Sikder and Hossain 2016; Mooney et al. 2017; Zeyaeyan et al. 2017; Hu et al. 2018; Jeworrek et al. 2019). Some studies found that Grell–Freitas performed best (Fowler et al. 2016; Sikder and Hossain 2016; Hu et al. 2018; Gao et al. 2017; Jeworrek et al. 2019) and outperformed Kain–Fritsch especially at higher resolutions (Sikder and Hossain 2016; Gao et al. 2017; Jeworrek et al. 2019). Other studies (Lim et al. 2014; Pennelly et al. 2014; Campos and Wang 2015; Stergiou et al. 2017; Ngailo et al. 2018) found Kain–Fritsch to be the better or best cumulus scheme.

Flaounas et al. (2011), Argüeso et al. (2011), and Klein et al. (2015) found that planetary boundary layer (PBL) parameterizations also have an impact on precipitation. Better performance was found using ACM2 (Argüeso et al. 2011; Ngailo et al. 2018), and/or YSU (Argüeso et al. 2011; Efstathiou et al. 2013).

Model sensitivity studies are often either limited in time (case studies, e.g., Zhang et al. 2017; Pu et al. 2019; Ngailo et al. 2018; Toride et al. 2019; Jeworrek et al. 2019) or constrained in the variation of model configurations (e.g., only one parameterization type is varied at a time; Jankov et al. 2005; Argüeso et al. 2011; Liu et al. 2011; Pennelly et al. 2014; Pérez et al. 2014; Meynadier et al. 2015; Cohen et al. 2015). This is because generating a long model dataset with a large number of different configurations is very computationally expensive.

Furthermore, model performance is usually sensitive to the region (Leutwyler et al. 2017; Mooney et al. 2017), and verification results are therefore mostly valid for the specific area over which the study was carried out. In the U.S. Pacific Northwest region, several studies [Colle et al. (1999, 2000), Colle and Mass (2000), Colle and Zeng (2004) and Garvert et al. (2005a,b), all using the mesoscale model MM5 (Grell et al. 1994)] observed overprediction of precipitation along the steep windward slopes and underprediction of precipitation in the lee of major barriers. Other studies found the opposite over interior western U.S. mountain ranges [Gowan et al. (2018) using the Weather Research and Forecasting (WRF) and the High-Resolution Rapid Refresh (HRRR) models] and also in the Pacific Northwest [Conrick and Mass (2019) and Darby et al. (2019) using the WRF, HRRR, and the Rapid Refresh (RAP) models].

Odon et al. (2019) showed that precipitation biases in reanalyses, which are based on coarser-resolution models, are mainly associated with deficiencies in terrain representation. Ralph et al. (2010) argued that excessive flow blocking can cause excessive upward forcing and precipitation upstream of barriers, and consequently overprediction of lee-side subsidence. Colle et al. (1999) suggested that the leeside dry bias could also result from the microphysics parameterizations neglecting horizontal advection of moisture, and generating/maintaining inadequate amounts of ice aloft.

Finer resolutions are often expected to improve forecasts because they can resolve smaller scales of terrain and surface features. For instance, intense convective precipitation forecasts appear to gain skill from increased spatial resolution (Roberts and Lean 2008; Roberts et al. 2009; Givati et al. 2012; Jang and Hong 2014). However, finer grids are more prone to the so-called double penalty problem (Rossa et al. 2008; Gilleland et al. 2009), where features are slightly shifted in time and/or space compared to the truth (as represented by observations or analysis), resulting in verification penalties in both space–time locations (Colle et al. 2000; Mass et al. 2002; Michaelides 2008).

When refining a model grid it is also important to be aware of the NWP gray zone (Zheng et al. 2016; Chow et al. 2019; Kealy 2019; Jeworrek et al. 2019): model setup can be challenging at grid spacings that are not fine enough to fully resolve processes explicitly, yet too fine to fully parameterize them

using approximating schemes. These “gray zone” scales differ for various processes (e.g., cumulus convection, turbulent eddies, orographic effects). New scale-aware schemes are increasingly developed to seamlessly bridge the gap between implicitly (i.e., parameterized) and explicitly (i.e., resolved) represented processes. For example, Gao et al. (2017) and Jeworrek et al. (2019) found the scale-aware Grell–Freitas cumulus parameterization to be most accurate across the convective gray zone in the United States.

The present study evaluates hourly precipitation forecasts from the WRF Model over the complex terrain of southwest BC. A selection of different parameterizations is systematically varied, including microphysics, cumulus, turbulence, and land surface schemes. Configurations are evaluated against station observations across different accumulation windows, forecast horizons, grid resolutions, and precipitation intensities.

This study differs from previous ones in that 1) it is a comprehensive evaluation over a full calendar year, and thus its results are more statistically robust than case studies, and 2) it comprises a large number of model configurations (>100), thoroughly exploring the available parameterization combinations. A wide variety of metrics are presented, with the intention that this study will contribute to the understanding of precipitation predictability using WRF across a range of applications, including disaster mitigation (e.g., floods, avalanches, debris flows) and optimization of clean energy (hydroelectric reservoir operations).

The methodology section contains an overview of model configurations, station observations, as well as the preparation of this dataset. The results section discusses individual model configuration performance, geographical and seasonal patterns, precipitation intensities, forecast horizons, and accumulation windows; and investigates similarities between the model groups. The last section summarizes the conclusions of this study.

2. Methodology

a. Modeling

This study uses the WRF Model (Skamarock et al. 2008) version 3.8.1 with the Advanced Research WRF (ARW) dynamical core to evaluate its performance with different parameterizations. The Global Deterministic Prediction System (GDPS) model (Côté et al. 1998; Girard et al. 2014) from Environment and Climate Change Canada (ECCC) downloaded on a $0.24^\circ \times 0.24^\circ$ grid provide initial conditions and 3-hourly boundary conditions. WRF runs are initialized daily at 0000 UTC for the year 2016, with a 3-day forecast horizon after each daily spinup. Model runs include a time-staggered spinup of 3 h for each nested subdomain, resulting in 9 h of total spinup that is excluded from the evaluation. We use an adaptive time step (Hutchinson 2007) to maintain numerical stability.

Fall 2016 was one of the warmest and wettest recorded on the South Coast of BC (Odon et al. 2017). Many parts of the region experienced a long-lasting and almost uninterrupted rain period resulting in an accumulated precipitation anomaly of over 200% at several locations (Odon et al. 2017). Managing

the unusual reservoir inflows was a challenge for the province's primary electric utility, BC Hydro.

Three two-way nested domains are employed with horizontal grid spacings of 27, 9, and 3 km to assess the dependence of the results on horizontal resolution (Fig. 1a), where the finer nest boundaries are separated from the parent domain boundaries by at least 20 grid points. The verification area for all grid sizes lies within the smallest domain, and spans several key hydroelectric reservoir watersheds in mountainous terrain. The WRF Model setup uses 65 sigma levels with a 50-hPa model top.

A systematic variation of three microphysics schemes, two cumulus convection schemes, two land surface models, and three combinations of PBL and surface layer schemes are tested (Table 1), yielding 36 different model configurations. Hereafter we use the abbreviations as specified in Table 1. These schemes were chosen because they are either commonly used and/or other studies showed sensitivity to their variation (see Introduction).

For instance, the YSU, ACM2, and GBM PBL parameterizations are tested in this study because of their proven wind speed forecast performance for wind farms in this region (Siuta et al. 2017). All selected PBL schemes work in connection with the same (and most popular) surface layer scheme: the updated MM5 similarity scheme (Dyer and Hicks 1970; Jiménez and Dudhia 2012).

The Unified Noah land surface model is a popular choice in WRF. In this study, it is compared to its newer version, Noah with multiparameterization options (Noah-MP, Niu et al. 2011). Among other updates, Noah-MP was redesigned to improve snow and land skin temperature, the diurnal cycle of soil temperature, and moisture, as well as the snowmelt runoff (Niu et al. 2011; Cai et al. 2014; Ma et al. 2017). These refinements resulted in reduced moisture and temperature biases (Duda et al. 2017; Wong et al. 2020) compared to the Rapid Update Cycle land surface model (Smirnova et al. 2016).

The longwave and shortwave radiation schemes are not varied: in our study we use RPTM (Mlawer et al. 1997) for longwave radiation and Dudhia (Dudhia 1989) for shortwave radiation. Several studies have found that radiation schemes have little impact on temperature and precipitation predictability (Fernández et al. 2007; Liu et al. 2011).

Models with grid spacings smaller than 4 km are often considered “convection permitting,” assuming that they are capable of resolving organized convection at this resolution (Weisman et al. 1997; Arakawa et al. 2011; Prein et al. 2015). However, such grid spacings are still insufficient to adequately trigger and represent small convective showers, individual convective cells, or updrafts (e.g., Bryan et al. 2003; Clark et al. 2016), and may require parameterized convection at least to some degree (e.g., Deng and Stauffer 2006; Lean et al. 2008; Roberts and Lean 2008). This gray zone problem was discussed in the introduction, and it is debatable whether the cumulus parameterization should be turned on in our finest 3-km domain. In this study we decided to include cumulus parameterizations in all domains and configurations, however, a conventional (KF) and a scale-aware (GF) option are used for comparison.

FIG. 1. (a) WRF Model domains. (b) The locations of 55 stations providing hourly precipitation records for forecast verification.

This study verifies raw model output without any post-processing. Postprocessing could affect the errors and ranking of the best performing models. However, the aim of this study is to understand the model behavior alone. Different post-processing techniques correct different error characteristics. For example, simple bias correction may reduce seasonal systematic errors, whereas more sophisticated techniques may reduce more complex conditional biases. A variety of verification metrics are shown in this study; users with a bias correction algorithm in place might put less weight on the systematic errors presented.

b. Verification

Hourly precipitation observations from 55 stations from two networks are used for verification: 26 stations from Environment and Climate Change Canada (ECCC) and 29 stations from BC Hydro (Fig. 1b). Station data from BC Hydro are used for the entire study period, whereas data from ECCC were available only for 7.5 out of the 12 months.

At lower elevations of BC's South Coast snowfall is rare. Thus, most observed precipitation fell as rain, and any amount of snow or sleet is represented as liquid equivalent. Snowfall measurement difficulties (undercatch, excessive evaporation from heated gauges, and delayed readings due to slow melting) may impair the observational dataset (Colle et al. 1999, 2000; Rasmussen et al. 2012).

Interpolation is needed to verify gridded model data against observations at point locations; however, studies show that interpolation does not have a large impact on verification results (e.g., Odon et al. 2019). The nearest-neighbor method compares the station observation with the closest model grid point. The present study adapts this technique, but adds another

step by averaging all observations that share the same model grid point as their nearest neighbor. This way, the same model grid point will not be verified multiple times in areas of dense observational networks, which would effectively give that model grid point more weight or importance than others. In the coarsest domain this reduces the 55 stations to 45 verification points, but in the finest domain each station has a unique nearest neighbor grid point.

Brief, intense precipitation rates are important to many users. However, verification of precipitation in hourly increments can raise double-penalty issues, where a timing error can, for example, result in an overforecast error in one hour and an underforecast in a subsequent hour. Extended accumulation windows can compensate for those timing errors and are useful to end users concerned with storm-total precipitation (e.g., BC Hydro), but they lose information about short-term rain intensities. In this verification study, a variety of different accumulation windows were investigated, however, the focus is on 6-hourly precipitation performance, which strikes a balance between capturing some short-duration precipitation rates while also allowing for some margin of temporal offset error. Furthermore, to limit the chance of discrete accumulation windows splitting a precipitation event and making it appear longer and less intense than it actually is, hourly rolling accumulation windows are used. This approach resamples the event in hourly steps to obtain at least one sample with optimal splitting considering the given time window.

The forecast skill of significant precipitation is typically more critical for decision-makers. Most hours have no precipitation, and most hourly precipitation totals are so small as to be insignificant to many forecast users. This study investigates the

TABLE 1. List of all of the tested WRF parameterizations (with abbreviations and references) that are varied in all possible combinations.

Physics parameterization scheme	Abbreviation	Reference
Cumulus convection		
Kain–Fritsch	KF	Kain (2004)
Grell–Freitas	GF	Grell and Freitas (2014)
Microphysics		
WRF Single-moment 5-class scheme	WSM5	Hong et al. (2004)
1.5-moment 6-class Thompson	Thom	Thompson et al. (2008)
2-moment 6-class Morrison	Morr	Morrison et al. (2009)
Planetary boundary layer		
Yonsei University scheme	YSU	Hong et al. (2006)
Asymmetric Convective Model	ACM2	Pleim (2007)
Grenier–Bretherton–McCaa scheme	GBM	Grenier and Bretherton (2001)
Land surface		
Unified Noah land surface model	Noah	Chen and Dudhia (2001)
Multiphysics Noah land surface model	Noah-MP	Niu et al. (2011); Yang et al. (2011)

forecast performance for the full spectrum of precipitation intensities, including an analysis of precipitation/no-precipitation events (where the measurable precipitation threshold is 0.25 mm), and some higher-impact subcategories such as the 75th percentile (“significant events”) and 95th percentile (“extreme events”). Since the precipitation climate varies across the domain, percentiles are calculated at each station based on observations, excluding 0-mm periods.

We consider several popular verification scores, including continuous and categorical metrics, the latter based on a contingency table. Details and equations beyond the following conceptual descriptions can be found in various references, such as Wilks (2011). Mean absolute error (MAE) is the average of the absolute differences between forecasts and observations. Bias is the mean difference between forecasts and observations, and can represent systematic error. The standard deviation (SD) of the error describes the spread of the differences between forecasts and observations and can serve as a measure of random error. The ratio of the SDs of model forecasts and observations [$SD(Fcst)/SD(Obs)$] compares the spread of the predicted and observed precipitation distributions, which are desired to be similar. Pearson correlation measures the strength of the linear relationship between forecasts and observations. The mean square difference (MSD) takes the average of the squared differences between forecasts and observations, giving more weight to larger errors than MAE. MSD can be decomposed into systematic and random error components (Willmott 1981).

For categorical forecasts, “accuracy” [also known as “proportion correct” (Wilks 2011)] indicates the proportion of forecasts that were correct. Correct forecasts include true negatives; hence, easily forecasted extended dry periods can increase accuracy. The false alarm ratio (FAR) is the fraction of incorrect forecasted positive events over the total number of forecasted positive events (different from the false alarm rate). Frequency bias describes the ratio of forecasted to observed positive events. Probability of detection (POD; also known as hit rate) is the fraction of correctly forecasted positive events

over observed positive events. The equitable threat score (ETS; also known as Gilbert skill score) is the “ratio of success” (Gilbert 1884) and takes into account the random chance of a hit.

3. Results and discussion

a. Individual model performance

Different model configurations perform best with respect to different error metrics; however, some configurations appear in the top rankings more often than others. Figure 2 gives an overview of various common verification metrics, while Fig. 3 shows categorical verification metrics derived from contingency table variables at two thresholds. For simplicity, metrics in both figures are shown only from the middle-resolution, 9-km grid, for 6-hourly rolling accumulation windows. All metrics are calculated from each time series and averaged over all stations. For 3- and 27-km grids (not shown), individual configuration rankings are similar to the 9-km grid, however, the values of the metrics differ.

From Fig. 2, the overarching findings are that configurations using Noah-MP are generally better than the ones using Noah, and KF is better than GF. GF configurations are competitive with the KF only when combined with YSU and Noah-MP. Microphysics–convective scheme combinations using KF perform best. WSM5–KF and Thom–KF have the best MAEs, whereas Thom–KF and Morr–KF have the best MSDs—the latter combinations therefore having fewer large errors.

In operational forecasting, raw model output typically undergoes postprocessing that improves the forecast performance. Bias is relatively easy to remove, whereas random error is more difficult. The ratio of random to total MSD error component indicates what portion of a configuration’s error may be more difficult to remove with postprocessing (Fig. 2). Namely, Thom–KF configurations, which have among the lowest MAEs and MSDs, are also likely to see the largest benefits from postprocessing because they have the best random/total MSD ratios and SD of errors.

MP	CU	PBL	LS	MAE	BIAS	SD(Error)	SD(Fcst) / SD(Obs)	Correlation	MSD	MSD random/total
WSM5	YSU	Noah		1.29	0.106	3.12	1.05	0.474	11.4	0.705
			N MP	1.27	0.0952	3.09	1.04	0.481	11.2	0.705
	KF	ACM2	Noah	1.29	0.109	3.16	1.07	0.476	11.8	0.719
			N MP	1.29	0.0905	3.15	1.06	0.477	11.7	0.712
	GBM	Noah		1.29	0.113	3.16	1.06	0.472	11.7	0.719
			N MP	1.25	0.0909	3.07	1.02	0.47	11.2	0.712
	YSU	Noah		1.31	0.113	3.2	1.09	0.469	12	0.727
			N MP	1.29	0.0922	3.14	1.06	0.477	11.6	0.714
	GF	ACM2	Noah	1.32	0.123	3.25	1.11	0.469	12.4	0.741
			N MP	1.31	0.12	3.23	1.1	0.472	12.3	0.739
	GBM	Noah		1.32	0.115	3.25	1.1	0.462	12.4	0.73
			N MP	1.31	0.109	3.24	1.1	0.467	12.4	0.736
Thom	YSU	Noah		1.28	0.117	3.07	1.02	0.475	11	0.671
			N MP	1.27	0.0988	3.05	1.01	0.475	10.8	0.669
	KF	ACM2	Noah	1.28	0.124	3.1	1.04	0.474	11.1	0.688
			N MP	1.26	0.1	3.04	1.02	0.483	10.8	0.68
	GBM	Noah		1.3	0.123	3.12	1.05	0.47	11.3	0.691
			N MP	1.28	0.106	3.09	1.04	0.478	11.1	0.689
	YSU	Noah		1.29	0.097	3.14	1.06	0.47	11.4	0.694
			N MP	1.27	0.0758	3.09	1.03	0.476	11.2	0.682
	GF	ACM2	Noah	1.31	0.126	3.21	1.1	0.463	11.9	0.716
			N MP	1.3	0.118	3.17	1.08	0.471	11.7	0.715
	GBM	Noah		1.31	0.114	3.19	1.08	0.46	11.7	0.713
			N MP	1.3	0.109	3.19	1.08	0.465	11.7	0.715
Morr	YSU	Noah		1.28	0.107	3.08	1.03	0.474	11	0.688
			N MP	1.28	0.0964	3.08	1.03	0.475	10.9	0.688
	KF	ACM2	Noah	1.29	0.107	3.12	1.05	0.47	11.3	0.701
			N MP	1.28	0.0963	3.11	1.04	0.474	11.3	0.699
	GBM	Noah		1.29	0.102	3.13	1.05	0.469	11.5	0.703
			N MP	1.29	0.0974	3.13	1.05	0.472	11.5	0.707
	YSU	Noah		1.31	0.138	3.17	1.08	0.469	11.5	0.71
			N MP	1.28	0.127	3.1	1.05	0.472	11.2	0.71
	GF	ACM2	Noah	1.32	0.151	3.21	1.1	0.466	11.9	0.731
			N MP	1.32	0.153	3.2	1.1	0.47	11.9	0.732
	GBM	Noah		1.33	0.141	3.24	1.11	0.458	12.1	0.73
			N MP	1.35	0.15	3.27	1.12	0.463	12.4	0.736

FIG. 2. Various location-averaged verification metrics for 6-hourly precipitation for each 9-km configuration. Dark blue colors are best, whereas dark red colors are worst. The color scale range is normalized to the range of ensemble (column) values, with white indicating values near the ensemble average. The perfect score for most metrics is zero, except for correlation and ratio of SDs of predicted and observed precipitation intensity distribution, which have a perfect score of one.

					0.25 mm					75 th Percentile				
					Accuracy	ETS	False Alarm Ratio	Frequency Bias	POD	Accuracy	ETS	False Alarm Ratio	Frequency Bias	POD
MP	CU	PBL	LS											
WSM5	YSU	Noah	N MP		0.813	0.338	0.412	1.17	0.67	0.922	0.218	0.614	1.14	0.415
					0.82	0.346	0.403	1.14	0.667	0.924	0.227	0.607	1.14	0.426
	KF	ACM2	Noah		0.816	0.34	0.404	1.14	0.659	0.923	0.219	0.615	1.13	0.414
					0.818	0.343	0.397	1.12	0.656	0.922	0.218	0.613	1.11	0.41
	GBM	Noah	N MP		0.816	0.34	0.411	1.16	0.666	0.923	0.22	0.614	1.13	0.416
					0.817	0.338	0.408	1.14	0.655	0.925	0.216	0.619	1.12	0.406
	YSU	Noah	N MP		0.817	0.341	0.396	1.1	0.651	0.92	0.217	0.616	1.14	0.414
					0.821	0.346	0.39	1.09	0.65	0.922	0.222	0.609	1.14	0.419
	GF	ACM2	Noah		0.82	0.339	0.395	1.09	0.643	0.921	0.217	0.62	1.16	0.417
					0.821	0.342	0.391	1.08	0.643	0.921	0.221	0.618	1.17	0.426
	GBM	Noah	N MP		0.818	0.338	0.396	1.09	0.641	0.921	0.216	0.62	1.15	0.413
					0.82	0.339	0.393	1.08	0.639	0.922	0.218	0.618	1.14	0.415
Thom	YSU	Noah	N MP		0.803	0.329	0.437	1.28	0.702	0.922	0.217	0.609	1.12	0.407
					0.808	0.336	0.426	1.25	0.697	0.924	0.22	0.601	1.1	0.407
	KF	ACM2	Noah		0.805	0.329	0.433	1.26	0.694	0.923	0.22	0.607	1.12	0.409
					0.809	0.336	0.424	1.23	0.69	0.924	0.224	0.601	1.11	0.414
	GBM	Noah	N MP		0.801	0.324	0.438	1.27	0.695	0.922	0.218	0.61	1.12	0.406
					0.807	0.334	0.425	1.24	0.691	0.923	0.223	0.602	1.11	0.411
	YSU	Noah	N MP		0.816	0.335	0.4	1.11	0.645	0.921	0.218	0.611	1.14	0.413
					0.818	0.34	0.396	1.1	0.647	0.922	0.221	0.606	1.13	0.414
	GF	ACM2	Noah		0.816	0.332	0.398	1.09	0.637	0.92	0.213	0.623	1.16	0.411
					0.818	0.336	0.395	1.09	0.638	0.921	0.218	0.618	1.17	0.419
	GBM	Noah	N MP		0.814	0.326	0.402	1.08	0.627	0.921	0.213	0.623	1.16	0.41
					0.815	0.33	0.396	1.07	0.629	0.921	0.218	0.618	1.16	0.417
Morr	YSU	Noah	N MP		0.809	0.331	0.423	1.2	0.676	0.922	0.217	0.613	1.13	0.41
					0.813	0.338	0.415	1.18	0.675	0.922	0.218	0.61	1.12	0.41
	KF	ACM2	Noah		0.812	0.332	0.418	1.18	0.666	0.923	0.216	0.613	1.12	0.408
					0.813	0.332	0.415	1.16	0.661	0.923	0.216	0.611	1.11	0.407
	GBM	Noah	N MP		0.81	0.33	0.421	1.19	0.668	0.922	0.217	0.613	1.12	0.408
					0.812	0.334	0.416	1.18	0.669	0.922	0.218	0.61	1.11	0.408
	YSU	Noah	N MP		0.812	0.332	0.414	1.16	0.661	0.921	0.216	0.615	1.16	0.415
					0.815	0.332	0.413	1.15	0.656	0.922	0.215	0.618	1.15	0.413
	GF	ACM2	Noah		0.812	0.325	0.414	1.15	0.649	0.921	0.214	0.622	1.18	0.417
					0.812	0.328	0.412	1.15	0.65	0.921	0.218	0.618	1.18	0.423
	GBM	Noah	N MP		0.811	0.322	0.415	1.13	0.641	0.921	0.212	0.623	1.16	0.409
					0.811	0.325	0.41	1.13	0.644	0.92	0.216	0.618	1.16	0.416

FIG. 3. As in Fig. 2, but for categorical metrics. Categories include precipitation/no-precipitation events (>0.25 mm) and 75th percentiles. Perfect scores: accuracy = 1; false alarm ratio = 0; frequency bias = 1; POD = 1; ETS = 1. Color interpretation is the same as for Fig. 2.

Looking at categorical, threshold-based metrics, clear performance differences exist between 0.25-mm thresholds (i.e., precipitation/no-precipitation) versus 75th percentiles (i.e., significant) events (Fig. 3). WSM5–KF, WSM5–GF, and Thom–GF perform best for >0.25 -mm events. For significant

(75th percentile) events, KF configurations have better accuracy, false alarm ratio, frequency bias, and ETS values than GF configurations. In particular, Thom–KF configurations do best. WSM5 configurations have better significant event POD. The 95th-percentile results (not shown) are similar to the 75th

FIG. 4. Relative monthly and annual precipitation bias for 27-km grids (colored shading, lines, and box-and-whisker plots; left axes) using (top) the GF cumulus configuration and (bottom) the KF cumulus configuration. Observed mean 6-hourly precipitation across all stations (gray bars; right axes).

percentile. Frequency biases show that all configurations produce rain more often than observed. Configurations with larger frequency bias values, such as the ones using Thom–KF at the >0.25 -mm threshold, consequently have higher false alarm ratios and lower accuracy values. However, because they are generally wetter, they also cause more hits, and hence have a higher/better POD in comparison. This pattern is often reversed between >0.25 -mm and significant/extreme events. For example, the same Thom–KF configurations show better frequency bias, false alarm ratio, and accuracy metrics at the 75th and 95th percentiles, but worse POD.

A Friedman omnibus test indicates significant differences among the configurations for all metrics and grid spacings (not shown) at the $\alpha = 0.05$ level. Analysis of pairwise differences among the configurations with the Nemenyi posthoc test shows that Thom–KF significantly outperforms almost all WSM5–GF and Morr–GF configurations for MSD and SD metrics, as well

as >0.25 -mm POD and 75th-percentile false alarm ratio. The differences between most Morr–GF and all other configurations are significant, especially concerning MAEs, MSDs, and SDs. Differences among the configurations are generally more significant for >0.25 -mm categorical verification metrics compared to 75th- and 95th-percentile scores. To highlight some configurations in particular, WSM5–KF–GBM–Noah–MP significantly outperforms almost all other configurations regarding MAEs and 75th-percentile accuracy, whereas Thom–KF–ACM2–NoahMP has significantly better correlation and 75th-percentile false alarm ratio than many other configurations.

b. Seasonal performance variation

As mentioned in the introduction, the local climatology primarily includes a cool, wet season and a warm, dry season. Precipitation is more stratiform and frontal in the cool season,

FIG. 5. Monthly and annual ranked MAE performance for 27-km grids using (top) the GF cumulus configurations and (bottom) the KF cumulus configurations. Ranks are out of all (GF and KF) configurations, but are separated into two plots for readability. Lower rank numbers mean better performance.

and more convective in the warm season. Moreover, the fall season 2016 was exceptionally warm and wet in the study area (see Introduction).

Monthly relative bias shows that all configurations generally exhibit a wet bias in the warm season and a neutral to slight dry bias in the cool season. This is most pronounced at the coarsest grid (Fig. 4). At finer grid spacings (not shown) monthly relative biases shift more toward negative (drier) values. Warm season wet bias is larger for configurations that use KF, whereas cool season bias is small and similar between the two cumulus schemes.

In the cool season, the best (lowest) ranks of 6-hourly precipitation MAE are mainly occupied by Thom–KF configurations (Fig. 5). However, in the more convective warm season,

the scale-aware GF configurations perform best. Some parameterization combinations have a larger performance variation with season than others. For example, those using Thom–KF switch from nearly consistently ranking best in the cool season to worst in the warm season, whereas those using WSM5–KF have a smaller variation in rank throughout the year.

The convective treatment has a large impact on summer precipitation performance. Although observed summer precipitation does not show a clear diurnal pattern (not shown), KF configurations exhibit elevated convective contributions to precipitation and false alarms in the afternoon, suggesting a causal relationship (Fig. 6). At 1400 LST, ~9.5% of all 9-km KF precipitation forecasts are false. The 27- and 3-km grids

FIG. 6. Diurnal variation of false alarm ratios (red) and ratio of convective parameterization contribution to total precipitation (gray), for hourly 9-km grid forecasts in summer (June–August).

show a similar pattern, with $\sim 12.5\%$ and $\sim 7.5\%$ false alarms, respectively. At the time of this peak, the KF parameterization contributes $\sim 40\%$ of the total precipitation for the 9-km grid. GF configurations do not produce this diurnal pattern and, hence, perform better overall in the warm season.

c. Geographical performance patterns

The previous sections compared station-averaged configuration performance, however, this varies widely across the diverse topography of the domain (see Introduction). Average MAEs range from 0.5 to $2.5 \text{ mm (6 h)}^{-1}$, where stations with wetter climatologies and lower predictability have higher MAEs. Pearson correlation coefficients range from 0.2 to 0.7 , with lower correlations over central Vancouver Island and over higher elevations of the Coast Range (not shown).

Relative biases across the region (Fig. 7) show generally widespread wet biases in summer, with no coherent geographical variations, which agrees with the summertime wet biases shown in the previous section (Figs. 4 and 6). The cool and wet season, which lasts from September throughout May, shows more coherent regional variations of relative bias (Fig. 7): most Vancouver Island stations have a dry bias, especially on the lee side of the terrain; Metro Vancouver and the Fraser Valley (see Fig. 1 for locations) exhibit a neutral to dry bias; and stations over the Coast Range and Northshore Mountains more often yield a wet bias.

While it may appear in Fig. 7 that all grids have similar bias distributions, the color scale does not extend to the true outliers. The extreme values in the bottom left corner of each subfigure reveal that some grid points have a very strong wet bias (150% – 250% more than climatological precipitation), especially at coarser grids, and that the bias in the cool season is

FIG. 7. Relative bias (WRF-Obs/Obs) of 6-hourly precipitation (unitless) by grid box as seasonal average plotted on model topography (in meters).

FIG. 8. 6-hourly precipitation error metrics for averages of different microphysics–cumulus scheme combinations and grid spacings. The arrows along the ordinates point toward values indicating better performance for each metric.

larger than in summer. These are mainly stations on the windward slopes of the Coast Range north of Vancouver. This finding could imply overdone orographic influences with underprediction (overprediction) on the leeward (windward) side of topographic barriers, as observed before by Colle et al. (1999, 2000) and Colle and Mass (2000) in the U.S. Pacific Northwest.

However, the windward slopes of the North Shore Mountains, immediately north of the Fraser Valley, are sometimes in the lee/rain shadow of the Vancouver Island and Olympic Mountains, depending on the prevailing flow and stability. Metro Vancouver, the Fraser Valley, and southern Vancouver Island show better performance (larger correlation coefficients, lower MAEs, and lower biases).

Some stations in the Coast Range show a cool-season wet bias. Rather than a modeling error, this could result from undercatchment of solid precipitation (Colle et al. 1999, 2000; Rasmussen et al. 2012), in particular at higher elevation stations where a greater percentage of precipitation falls as snow in the cool season.

Figure 7 also shows the significant differences in model terrain for each grid. For example, the highest mountain in the verification area (Mount Waddington, see Fig. 1b) has a

measured elevation of 4019 m, but is represented by approximately 2000-m (coarsest) to 2900-m (finest) gridbox elevations. Vertical relief is further diluted by nearby valleys that are not low enough. Coarser grid spacings do not contain finer-scale terrain detail and amplitude, and therefore lack the resulting impacts to flow.

d. Resolution-dependent performance

We use two-way nesting; hence, all three resolutions, and the scales and processes represented by them, all interact. Finer grids have worse MAE and MSD performance (Figs. 8a,b). Surprisingly, the grid dependency of MAEs is most apparent in configurations using GF, even though this cumulus parameterization uses a scale-aware mass-flux approach that should improve convective triggering at higher resolutions. MAEs are especially large when using Morr–GF at 9 and 3 km.

Pearson correlation coefficients generally worsen (decrease) with finer grid spacing (Fig. 8d). While the change with resolution is larger than the change among the scheme combinations, KF configurations have slightly better performance. All correlation coefficients lie approximately between 0.45 and 0.5, which indicates a fairly weak linear

FIG. 9. Histograms of observed precipitation percentile thresholds across all stations for different accumulation windows (55 stations total).

association between the observations and forecasts from all models.

Coarser grids have a worse (higher) wet relative bias on average (Fig. 8c). However, bias is sign sensitive, hence, dry biases may exist for some configurations, grid points, or times of the year, but they can be cancelled out by larger wet biases.

The spread of the predicted model precipitation intensity distribution is on average wider than observed, more so for coarser grids and GF configurations (Fig. 8e). However, the error SD is worse (larger) at finer grids on average, which is why MAEs are larger at finer grids despite their reduced bias (Fig. 8f). This may be expected given that finer grids tend to produce more local and extreme values in their representation of finer spatial scales. However, minor displacement of these small-scale features results in outsized errors (i.e., double penalties), and larger SDs.

e. Common versus extreme event performance

There are a variety of precipitation forecast applications that value different characteristics of a forecast. For example, some might prioritize good discrimination between precipitation/no precipitation, while others prioritize accurate heavy precipitation forecasts. To give the reader a sense of the disparate climatologies of the stations across the region, the distribution of observed 75th- and 95th-percentile thresholds are plotted in Fig. 9.

Looking at hits, misses, and false alarms across different precipitation intensities, the difference among the individual models is relatively small in comparison to differences over accumulation windows and resolutions. The mean of all configurations at each resolution is shown in Fig. 10. Correct rejections are not included in Fig. 10, because they are not as informative for precipitation because they are easily achieved (as most of the time it is not precipitating), even more so for 75th- and 95th-percentile thresholds. However, the proportion of correct-negative events can be estimated from the scale of Fig. 10 since it displays the fractions of all events.

Accumulation window (i.e., temporal resolution) has a larger impact on forecast performance than the grid spacing (i.e., spatial resolution). Hit rate decreases for more extreme events, smaller accumulation time windows, and finer grid resolutions, which is expected as these are more difficult forecasts. This agrees with ETS values (Fig. 11), which also significantly improve with longer accumulation windows, and are often slightly better at the 27-km grid spacing.

The general overprediction of precipitation frequency is reflected in a false-alarm rate that often exceeds the miss rate, especially on the coarsest grid (Fig. 10). Finer-resolution miss rates are slightly worse (larger), whereas coarser-resolution false alarm rates are worse (larger). As accumulation windows get shorter and grid spacing gets finer, correct rejections (not shown) improve (increase). Accordingly, the total number of all correct forecasts (viz., true positives, i.e., hits, plus true negatives) increases with finer grids and shorter accumulation windows.

This study finds that finer spatial resolutions score better in metrics that give credit for correct rejections. However, for forecasts that exclude these easily achieved scores, the ratio of successful positive forecasts (hits only) to unsuccessful forecasts (misses plus false alarms) diminishes rapidly with smaller accumulation windows and more extreme events. This means that the “deterministic limit” (Hewson 2020) is exceeded and a positive forecast of such an event is on average more likely to be incorrect than correct.

f. Predictability with forecast horizon and accumulation window

The diminishing forecast quality with longer forecast horizons and shorter accumulation windows is reflected in an increase in 1-h equivalent (normalized by time period) MAE and a decrease in correlation (Fig. 12). Six-hourly MAE increases on average by 6% from day 1 to day 2, and by 3% from day 2 to day 3 (not shown). However, no matter how long the accumulation window, Fig. 12 shows that day-1 forecast performance is best, and the difference between day

FIG. 10. The ratio of hits, misses, and false alarms for different accumulation windows, grid spacings, and precipitation intensities for the mean of all configurations.

1 and day 2 forecasts remains larger than that between day 2 and day 3.

Ensemble-mean MAE and correlation change asymptotically with extended accumulation window (Fig. 12). After normalizing by accumulation window length, the 1-h equivalent MAE improves (reduces) by over 25% from hourly to 12-hourly accumulation windows and by about 50% from hourly to daily accumulation intervals. The improvement levels out after about 2 or 3 days of accumulation. The finest grid exhibits the worst MAEs irrespective of accumulation period and forecast day. This illustrates that temporal averaging benefits MAEs

at all grid spacings similarly and, at finer grids, it cannot compensate for the larger random error that results from the enhanced spatial detail. The differences between the 3- and 9-km grids are consistently small, while the 27-km grid is consistently best, although the difference among the domains becomes smaller with increasing accumulation window.

Pearson correlation coefficient also improves rapidly when extending the accumulation within the first day, and levels out after 1–2 days. However, the coarsest grid is only best for accumulation periods up to 1 day, at which point the finer grids (especially the 9-km grid) become better.

FIG. 11. ETS values for different accumulation windows, grids, and precipitation intensities for the mean of all configurations. Error bars represent the spread among the individual models. Higher ETS is better.

FIG. 12. 1-h-equivalent MAE (thin lines corresponding to the left axis) and Pearson correlation coefficients (thick lines corresponding to the right axis) for increasing accumulation windows up to 7 days. [For example, for a day-1 forecast horizon (dark blue), 96-hourly (4-day) scores are based on the summation of precipitation from forecast hours 1 to 24 from four consecutive forecast initializations—it is a theoretical 4-day accumulation. This calculation is done using a rolling four-initialization window across the whole forecast dataset period.]

Longer accumulation windows are more likely to capture the entirety of a precipitation event and compensate for potential timing errors between forecasted and observed precipitation. On the other hand, important information about variable precipitation rates at time scales shorter than a given accumulation window are averaged out. Knowledge of both performance characteristics are important because shorter-duration precipitation intensities may be more important to some users, whereas storm total precipitation may be more important to others.

g. Model interdependence

For different output variables, some parameterization types will impact model results more than others: e.g., precipitation forecasts are expected to be primarily affected by the microphysics and cumulus parameterization. But which group of parameterizations has the largest impact on precipitation forecast performance, and how much do the choices of PBL and land surface parameterizations play a role? Here, hierarchical clustering (based on Euclidean distance and average linkage) is used to group the numerous models based on their Pearson correlation coefficients. The resulting heat map with corresponding dendrograms are shown in Figs. 13 and 14, for the coarsest and finest grids, respectively. This type of analysis is useful for insights into scheme performance, but also for constructing a diverse multiphysics ensemble. That is, choosing configurations that belong to unrelated cluster groups will make for a more diverse NWP ensemble, which is a desired characteristic for ensemble members that are nearly equally skillful (Eckel and Mass 2005; Lee et al. 2012; Krishnamurti et al. 2016).

All configurations are highly correlated with each other, as Pearson correlation coefficients are generally large, especially in the 27-km domain (>0.9 ; Fig. 13). The configurations that use KF produce very similar precipitation, whereas GF models are less correlated. The next-level clustering is mainly based on the microphysics, where Thom and Morr are better correlated while WSM5 remains in its own subcluster. PBL choice, however, also plays a secondary role, especially within the GF group (where Morr and Thom in combination with GBM and ACM2 are grouped together).

The 3-km domains (Fig. 14) are slightly less (yet still well) correlated, with coefficients >0.8 . Their first-level clustering is based on a mix of microphysics and cumulus parameterizations and groups into configurations using 1) Thom–KF and Thom–GF, 2) Morr–KF and WSM5–KF, and 3) Morr–GF and WSM5–GF.

Configuration clustering differs slightly among the three resolutions because parameterized subgrid processes can be scale-dependent (e.g., the gray zone). Clustering in the 9-km grid (not shown) is more similar to the one from the 27-km grid (Fig. 13).

The choice of PBL and land surface parameterizations are of secondary importance for precipitation, but ACM2 and GBM are often grouped together, while YSU remains in its own subgroup. The choice of land surface scheme is the least decisive factor. Different clustering techniques and forecast horizons (not shown) yield similar results.

A χ^2 test was conducted to pairwise compare the forecasted categorical precipitation intensity distributions using the following nonoverlapping bins: [0, 0.25], (0.25, 1.0], (1.0, 2.5], (2.5, 5.0], (5.0, 10.0], (10.0, 20.0] and (20.0, ∞). The “X” marks in Figs. 13 and 14 mark configuration pairs for which the null

FIG. 13. Heat map of Pearson correlation coefficients resulting from hierarchical clustering of all configurations for the coarsest domain. First- or second-level clusters (see dendrograms) are framed in black to highlight groups of configurations. Configuration pairs *without* “X” marks have statistically significant differences in precipitation intensity distribution.

hypothesis of homogeneity could not be rejected at the Bonferroni-corrected $\alpha = 0.05/630 = 7.94 \times 10^{-5}$ significance level. That is, the χ^2 test shows that precipitation intensity distributions differ significantly among most configurations.

4. Summary and conclusions

A systematic variation of four parameterization types (microphysics, cumulus, PBL, and land surface), resulting in over 100 configurations of the WRF Model, were evaluated over a full year of 3-day forecasts, initialized once per day. Both continuous and categorical statistics were used to verify precipitation over the complex terrain of southwest British Columbia, Canada, across different resolutions, accumulation windows, seasons, and precipitation intensities.

Cumulus and microphysics parameterizations together produce the total precipitation in these model configurations (and generally in most NWP models), and this study confirms that they are the parameterizations that primarily determine precipitation forecast performance. PBL parameterizations have secondary importance, and land surface parameterizations had the least impact on precipitation forecast performance.

Slight, yet consistent, improvements were seen when using Noah-MP instead of the older Noah land surface model. This is likely an indirect result of the reduced diurnal temperature bias that was observed in Noah-MP as compared to Noah (not shown).

No consistent performance improvement was seen for any individual PBL parameterization, but they had an impact on precipitation dependent on their combination with the

FIG. 14. As in Fig. 13, but for the finest domain.

microphysics and cumulus parameterizations. Another important consideration to operational forecasting is that YSU was computationally the fastest PBL parameterization (on average 11% faster than ACM2 and GBM, where ACM2 is marginally faster than GBM).

The best performing microphysical parameterizations were either WSM5 or Thom. However, the WSM5 was the most computationally inexpensive, and our verification shows that it yielded competitive verification scores compared to more sophisticated and computationally expensive parameterizations. Thom and Morr took on average 20% longer to run than WSM5, with Thom being marginally faster than Morr. This agrees with other studies that concluded that the most complex parameterizations are not needed to produce good precipitation forecasts (Colle and Mass 2000; Zeyaeyan et al. 2017; Hu et al. 2018; Conrick and Mass 2019; Jeworrek et al. 2019).

The scale-aware GF cumulus parameterization did not outperform the conventional KF scheme at finer resolutions, contrary to what one might expect (Gao et al. 2017; Kwon and Hong 2017; Jeworrek et al. 2019). Although GF performed better for summertime convective precipitation, when the conventional KF parameterization produced an unrealistic diurnal pattern, KF performed better across all scales for wintertime frontal precipitation, which contributes to the majority of the annual rainfall in our study region of southwest BC.

The “best-performing” configuration is unique to each user and their application, and accordingly so is the importance of different verification metric(s). However, in southwest BC the following five models arguably perform better than average across most metrics (in no specific order):

- WSM5-KF-YSU-NoahMP,
- WSM5-KF-GBM-NoahMP,

- Thom–KF–YSU–NoahMP,
- Thom–KF–ACM2–NoahMP,
- Thom–GF–YSU–NoahMP.

During its cool season southwest BC receives large amounts of orographically enhanced frontal precipitation. Some grid points located on windward slopes or at higher elevations showed particularly strong cool-season wet biases, especially for coarser grids, whereas grid points in the lee of the Vancouver Island mountains had distinct dry biases. This suggests orographic influences are overdone in most configurations. Fortunately for the public interest, more densely populated areas such as metro Vancouver, the Fraser Valley, and southern Vancouver Island showed overall good verification scores year round. In the warm season, which is drier, the relative biases were generally smaller.

Coarser resolutions, which effectively spatially average precipitation, had smaller random errors, yield smaller MAEs, and higher correlation coefficients compared to the finer grids. Although coarser grids suffered from larger biases (systematic errors), these can be largely and easily reduced using bias correction, which would greatly improve the 27-km forecasts. Higher resolutions had higher random errors, likely because they are affected by double-penalty issues that can dominate and distort verification. However, they have a more realistic spread within their total precipitation intensity distributions and contain valuable spatial, temporal, and intensity information that coarser resolutions are unable to represent. Finer grids performed best for metrics like bias, frequency bias, and categorical-forecast equitable threat scores and accuracy. These metrics indicate that finer grids had the largest fraction of correct forecasts, when including correct rejections (true negatives; which are often the majority). Significant or extreme event performance matters most for many users. The 27-km grid had the highest relative hit rate for these events.

When conducting and interpreting a verification study comparing different resolutions, our results show that it is important to carefully consider the verification techniques and preprocessing of the dataset to grant a fair comparison (e.g., grid spacing, temporal averaging, and station interpolation). For example, the length of accumulation windows had a larger impact on forecast performance than different grid spacings, and different configurations.

The various coarser-resolution configurations were highly correlated with each other, especially among configurations that used KF. Configurations with finer grids diverged more, and the choice of cumulus and microphysics parameterization combinations became increasingly important. This is expected because the coarsest WRF domains are more similar, as they receive their input from the same initial and boundary conditions, while successive nested domains are less directly influenced by them. Last, the present study yielded high correlations among PBL parameterizations, indicating little effect on the precipitation forecast on average.

The results of this verification study should be considered under these limitations, among others: 1) it was done for one year, and each year is unique, depending on, e.g., the phase of

climatological indices; 2) it used station observations, which are subject to observational errors, and compared point observations (with limited spatial coverage over the region) to a gridded forecast; 3) it was done over southwest BC, so while results may have some application to climatologically similar regions, model performance likely varies across climatologically disparate regions; 4) although it covered numerous combinations of physics parameterizations, it was not feasible to try all combinations; 5) it used two-way nesting in which all domains affect each other; 6) it intentionally used raw model output, whereas postprocessing bias correction is common practice in operational forecasts; and 7) it used only one initial and boundary condition each day because the focus was on parameterization combinations.

Despite the inevitable limitations evaluation studies entail, the present study was relatively rigorous and unique in that it systematically varied model configurations across many observing sites for a full year. We feel it provides valuable information about the predictability of precipitation in WRF across a variety of metrics and methods, particularly in wet coastal regions with steep mountainous terrain. We hope that our findings can help WRF users and forecasters with their model configurations, ensemble compositions, and resulting interpretation; and inform WRF developers about the performance characteristics of the WRF Model and its parameterizations.

Acknowledgments. The computational and storage resources required to generate and manage the extensive model dataset were provided by WestGrid and Compute Canada (www.computecanada.ca) through the Resource Allocation Competition awards 2018 and 2019. The research was enabled by funding support provided by Mitacs (Grant IT07224), BC Hydro (Grant C0567290), the Natural Science and Engineering Research Council (NSERC; Grant RGPIN-2017-03849), and the University of British Columbia (UBC) Geophysical Disaster Computational Fluid Dynamics Center. We also thank Timothy Chun-Yiu Chui, Yingkai Sha, Pedro Odon, Henryk Modzelewski, and Roland Schigas for their support with this study.

Data availability statement. The model dataset analyzed in this study is too large to be archived. Model point forecasts are not publicly archived, but can be made available upon request [contact Roland Stull (rstull@eoas.ubc.ca)]. ECCC station data used for verification are available at https://climate.weather.gc.ca/historical_data/search_historic_data_e.html, whereas BC Hydro station data may be obtained by contacting Gregory West (greg.west@bchydro.com).

REFERENCES

- Arakawa, A., J.-H. Jung, and C.-M. Wu, 2011: Toward unification of the multiscale modeling of the atmosphere. *Atmos. Chem. Phys.*, **11**, 3731–3742, <https://doi.org/10.5194/acp-11-3731-2011>.
- Argüeso, D., J. M. Hidalgo-Muñoz, S. R. Gámiz-Fortis, M. J. Esteban-Parra, J. Dudhia, and Y. Castro-Díez, 2011: Evaluation of WRF parameterizations for climate studies over southern Spain using a multistep regionalization. *J. Climate*, **24**, 5633–5651, <https://doi.org/10.1175/JCLI-D-11-00073.1>.

- Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.*, **131**, 2394–2416, [https://doi.org/10.1175/1520-0493\(2003\)131<2394:RRFTSO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2).
- Cai, X., Z. L. Yang, C. H. David, G. Y. Niu, and M. Rodell, 2014: Hydrological evaluation of the noah-MP land surface model for the Mississippi River basin. *J. Geophys. Res. Atmos.*, **119**, 23–38, <https://doi.org/10.1002/2013JD020792>.
- Campos, E., and J. Wang, 2015: Numerical simulation and analysis of the April 2013 Chicago floods. *J. Hydrol.*, **531**, 454–474, <https://doi.org/10.1016/j.jhydrol.2015.09.004>.
- Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part II: Preliminary model validation. *Mon. Wea. Rev.*, **129**, 587–604, [https://doi.org/10.1175/1520-0493\(2001\)129<0587:CAALSH>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0587:CAALSH>2.0.CO;2).
- Chow, F. K., C. Schär, N. Ban, K. A. Lundquist, L. Schlemmer, and X. Shi, 2019: Crossing multiple gray zones in the transition from mesoscale to microscale simulation over complex terrain. *Atmosphere*, **10**, 274, <https://doi.org/10.3390/atmos10050274>.
- Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, <https://doi.org/10.1002/met.1538>.
- Cohen, A. E., S. M. Cavallo, M. C. Coniglio, H. E. Brooks, A. E. Cohen, S. M. Cavallo, M. C. Coniglio, and H. E. Brooks, 2015: A review of planetary boundary layer parameterization schemes and their sensitivity in simulating southeastern U.S. cold season severe weather environments. *Wea. Forecasting*, **30**, 591–612, <https://doi.org/10.1175/WAF-D-14-00105.1>.
- Colle, B. A., and C. F. Mass, 2000: The 5–9 February 1996 flooding event over the Pacific Northwest: Sensitivity studies and evaluation of the MM5 precipitation forecasts. *Mon. Wea. Rev.*, **128**, 593–617, [https://doi.org/10.1175/1520-0493\(2000\)128<0593:TFFEOT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128<0593:TFFEOT>2.0.CO;2).
- , and Y. Zeng, 2004: Bulk microphysical sensitivities within the MM5 for orographic precipitation. Part I: The Sierra 1986 event. *Mon. Wea. Rev.*, **132**, 2780–2801, <https://doi.org/10.1175/MWR2821.1>.
- , K. J. Westrick, and C. F. Mass, 1999: Evaluation of MM5 and Eta-10 precipitation forecasts over the Pacific Northwest during the cool season. *Wea. Forecasting*, **14**, 137–154, [https://doi.org/10.1175/1520-0434\(1999\)014<0137:EOMAEP>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0137:EOMAEP>2.0.CO;2).
- , C. F. Mass, and K. J. Westrick, 2000: MM5 precipitation verification over the Pacific Northwest during the 1997–99 cool seasons. *Wea. Forecasting*, **15**, 730–744, [https://doi.org/10.1175/1520-0434\(2000\)015<0730:MPVOTP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0730:MPVOTP>2.0.CO;2).
- , J. B. Wolfe, J. W. Steenburgh, D. E. Kingsmill, J. A. Cox, and J. C. Shafer, 2005: High-resolution simulations and microphysical validation of an orographic precipitation event over the Wasatch Mountains during IPEX IOP3. *Mon. Wea. Rev.*, **133**, 2947–2971, <https://doi.org/10.1175/MWR3017.1>.
- Conrick, R., and C. F. Mass, 2019: An evaluation of simulated precipitation characteristics during OLYMPEx. *J. Hydrometeorol.*, **20**, 1147–1164, <https://doi.org/10.1175/JHM-D-18-0144.1>.
- Cookson-Hills, P., D. J. Kirshbaum, M. Surcel, J. G. Doyle, L. Fillion, D. Jacques, and S. J. Baek, 2017: Verification of 24-h quantitative precipitation forecasts over the Pacific northwest from a high-resolution ensemble Kalman filter system. *Wea. Forecasting*, **32**, 1185–1208, <https://doi.org/10.1175/WAF-D-16-0180.1>.
- Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998: The operational CMC-MRB global environmental multi-scale (GEM) model. Part I: Design considerations and formulation. *Mon. Wea. Rev.*, **126**, 1373–1395, [https://doi.org/10.1175/1520-0493\(1998\)126<1373:TOCMGE>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<1373:TOCMGE>2.0.CO;2).
- Courant, R., K. Friedrichs, and H. Lewy, 1928: Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, **100**, 32–74, <https://doi.org/10.1007/BF01448839>.
- Darby, L. S., A. B. White, D. J. Gottas, and T. Coleman, 2019: An evaluation of integrated water vapor, wind, and precipitation forecasts using water vapor flux observations in the western United States. *Wea. Forecasting*, **34**, 1867–1888, <https://doi.org/10.1175/WAF-D-18-0159.1>.
- Deng, A., and D. R. Stauffer, 2006: On improving 4-km mesoscale model simulations. *J. Appl. Meteor. Climatol.*, **45**, 361–381, <https://doi.org/10.1175/JAM2341.1>.
- Derin, Y., and K. K. Yilmaz, 2014: Evaluation of multiple satellite-based precipitation products over complex topography. *J. Hydrometeorol.*, **15**, 1498–1516, <https://doi.org/10.1175/JHM-D-13-0191.1>.
- Di Luca, A., E. Flaounas, P. Drobinski, and C. L. Brossier, 2014: The atmospheric component of the Mediterranean Sea water budget in a WRF multi-physics ensemble and observations. *Climate Dyn.*, **43**, 2349–2375, <https://doi.org/10.1007/s00382-014-2058-z>.
- Duda, J. D., X. Wang, and M. Xue, 2017: Sensitivity of convection-allowing forecasts to land surface model perturbations and implications for ensemble design. *Mon. Wea. Rev.*, **145**, 2001–2025, <https://doi.org/10.1175/MWR-D-16-0349.1>.
- Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107, [https://doi.org/10.1175/1520-0469\(1989\)046<3077:NSOCOD>2.0.CO;2](https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2).
- Dyer, A. J., and B. B. Hicks, 1970: Flux-gradient relationships in the constant flux layer. *Quart. J. Roy. Meteor. Soc.*, **96**, 715–721, <https://doi.org/10.1002/qj.49709641012>.
- Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, <https://doi.org/10.1175/WAF843.1>.
- Efstathiou, G. A., N. M. Zoumakis, D. Melas, C. J. Lolis, and P. Kassomenos, 2013: Sensitivity of WRF to boundary layer parameterizations in simulating a heavy rainfall event using different microphysical schemes. Effect on large-scale processes. *Atmos. Res.*, **132–133**, 125–143, <https://doi.org/10.1016/j.atmosres.2013.05.004>.
- Fan, X., 2009: Impacts of soil heating condition on precipitation simulations in the weather research and forecasting model. *Mon. Wea. Rev.*, **137**, 2263–2285, <https://doi.org/10.1175/2009MWR2684.1>.
- Fernández, J., J. P. Montavez, J. Saenz, J. F. Gonzalez-Rouco, and E. Zorita, 2007: Sensitivity of the MM5 mesoscale model to physical parameterizations for regional climate studies: Annual cycle. *J. Geophys. Res.*, **112**, D04101, <https://doi.org/10.1029/2005JD006649>.
- Flaounas, E., S. Bastin, and S. Janicot, 2011: Regional climate modelling of the 2006 West African monsoon: Sensitivity to convection and planetary boundary layer parameterisation using WRF. *Climate Dyn.*, **36**, 1083–1105, <https://doi.org/10.1007/s00382-010-0785-3>.
- Fowler, L. D., W. C. Skamarock, G. A. Grell, S. R. Freitas, and M. G. Duda, 2016: Analyzing the Grell–Freitas convection scheme from hydrostatic to nonhydrostatic scales within a global model. *Mon. Wea. Rev.*, **144**, 2285–2306, <https://doi.org/10.1175/MWR-D-15-0311.1>.
- Gao, Y., L. R. Leung, C. Zhao, and S. Hagos, 2017: Sensitivity of U.S. summer precipitation to model resolution and con-

- vective parameterizations across gray zone resolutions. *J. Geophys. Res. Atmos.*, **122**, 2714–2733, <https://doi.org/10.1002/2016JD025896>.
- García-Díez, M., J. Fernández, and R. Vautard, 2015: An RCM multi-physics ensemble over Europe: Multi-variable evaluation to avoid error compensation. *Climate Dyn.*, **45**, 3141–3156, <https://doi.org/10.1007/s00382-015-2529-x>.
- Garvert, M. F., B. A. Colle, and C. F. Mass, 2005a: The 13–14 December 2001 IMPROVE-2 event. Part I: Synoptic and mesoscale evolution and comparison with a mesoscale model simulation. *J. Atmos. Sci.*, **62**, 3474–3492, <https://doi.org/10.1175/JAS3549.1>.
- , C. P. Woods, B. A. Colle, C. F. Mass, P. V. Hobbs, M. T. Stoelinga, and J. B. Wolfe, 2005b: The 13–14 December 2001 IMPROVE-2 event. Part II: Comparisons of MM5 model simulations of clouds and precipitation with observations. *J. Atmos. Sci.*, **62**, 3520–3534, <https://doi.org/10.1175/JAS3551.1>.
- Gilbert, G. K., 1884: Finley's tornado predictions. *Amer. Meteor. J.*, **1**, 166–172.
- Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, <https://doi.org/10.1175/2009WAF2222269.1>.
- Girard, C., and Coauthors, 2014: Staggered vertical discretization of the Canadian Environmental Multiscale (GEM) model using a coordinate of the log-hydrostatic-pressure type. *Mon. Wea. Rev.*, **142**, 1183–1196, <https://doi.org/10.1175/MWR-D-13-00255.1>.
- Givati, A., B. Lynn, Y. Liu, and A. Rimmer, 2012: Using the WRF Model in an operational streamflow forecast system for the Jordan River. *J. Appl. Meteor. Climatol.*, **51**, 285–299, <https://doi.org/10.1175/JAMC-D-11-082.1>.
- Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, <https://doi.org/10.1175/WAF-D-17-0144.1>.
- Grell, G. A., and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, **14**, 5233–5250, <https://doi.org/10.5194/acp-14-5233-2014>.
- , J. Dudhia, and D. Stauffer, 1994: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note NCAR/TN-398+STR, 121 pp., <https://doi.org/10.5065/D60Z716B>.
- Grenier, H., and C. S. Bretherton, 2001: A moist PBL parameterization for large-scale models and its application to subtropical cloud-topped marine boundary layers. *Mon. Wea. Rev.*, **129**, 357–377, [https://doi.org/10.1175/1520-0493\(2001\)129<0357:AMPPFL>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0357:AMPPFL>2.0.CO;2).
- Hewson, T., 2020: New approaches to verifying forecasts of hazardous weather. Accessed 20 July 2020, <https://www.cawcr.gov.au/projects/verification/Hewson/DeterministicLimit.html>.
- Hong, S.-Y., J. Dudhia, and S.-H. Chen, 2004: A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Mon. Wea. Rev.*, **132**, 103–120, [https://doi.org/10.1175/1520-0493\(2004\)132<0103:ARATIM>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<0103:ARATIM>2.0.CO;2).
- , Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, <https://doi.org/10.1175/MWR3199.1>.
- Hu, X.-M., M. Xue, R. A. McPherson, E. Martin, D. H. Rosendahl, and L. Qiao, 2018: Precipitation dynamical downscaling over the Great Plains. *J. Adv. Model. Earth Syst.*, **10**, 421–447, <https://doi.org/10.1002/2017MS001154>.
- Hutchinson, T. A., 2007: An adaptive time-step for increased model efficiency. *Eighth WRF Users' Workshop*, Boulder, CO, NCAR, 9.4, http://www2.mmm.ucar.edu/wrf/users/workshops/WS2007/abstracts/9-4_Hutchinson.pdf.
- Jang, J., and S.-Y. Hong, 2014: Quantitative forecast experiment of a heavy rainfall event over Korea in a global model: Horizontal resolution versus lead time issues. *Meteor. Atmos. Phys.*, **124**, 113–127, <https://doi.org/10.1007/s00703-014-0312-x>.
- Jankov, I., W. A. Gallus, M. Segal, B. Shaw, and S. E. Koch, 2005: The impact of different WRF Model physical parameterizations and their interactions on warm season MCS rainfall. *Wea. Forecasting*, **20**, 1048–1060, <https://doi.org/10.1175/WAF888.1>.
- Jeworrek, J., G. West, and R. Stull, 2019: Evaluation of cumulus and microphysics parameterizations in WRF across the convective gray zone. *Wea. Forecasting*, **34**, 1097–1115, <https://doi.org/10.1175/WAF-D-18-0178.1>.
- Jiménez, P. A., and J. Dudhia, 2012: Improving the representation of resolved and unresolved topographic effects on surface wind in the WRF Model. *J. Appl. Meteor. Climatol.*, **51**, 300–316, <https://doi.org/10.1175/JAMC-D-11-084.1>.
- Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.*, **43**, 170–181, [https://doi.org/10.1175/1520-0450\(2004\)043<0170:TKCPAU>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2).
- Kealy, J. C., 2019: Probing the ‘grey zone’ of NWP – Is higher resolution always better? *Weather*, **74**, 246–249, <https://doi.org/10.1002/wea.3506>.
- Klein, C., D. Heinzeller, J. Bliefernicht, and H. Kunstmann, 2015: Variability of West African monsoon patterns generated by a WRF multi-physics ensemble. *Climate Dyn.*, **45**, 2733–2755, <https://doi.org/10.1007/s00382-015-2505-5>.
- Klemp, J. B., 2011: A terrain-following coordinate with smoothed coordinate surfaces. *Mon. Wea. Rev.*, **139**, 2163–2169, <https://doi.org/10.1175/MWR-D-10-05046.1>.
- Krishnamurti, T. N., V. Kumar, A. Simon, A. Bhardwaj, T. Ghosh, and R. Ross, 2016: A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Rev. Geophys.*, **54**, 336–377, <https://doi.org/10.1002/2015RG000513>.
- Kwon, Y. C., and S.-Y. Hong, 2017: A mass-flux cumulus parameterization scheme across gray-zone resolutions. *Mon. Wea. Rev.*, **145**, 583–598, <https://doi.org/10.1175/MWR-D-16-0034.1>.
- Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office unified model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, **136**, 3408–3424, <https://doi.org/10.1175/2008MWR2332.1>.
- Lee, J. A., W. C. Kolczynski, T. C. McCandless, and S. E. Haupt, 2012: An objective methodology for configuring and down-selecting an NWP ensemble for low-level wind prediction. *Mon. Wea. Rev.*, **140**, 2270–2286, <https://doi.org/10.1175/MWR-D-11-00065.1>.
- Leutwyler, D., D. Lüthi, N. Ban, O. Fuhrer, and C. Schär, 2017: Evaluation of the convection-resolving climate modeling approach on continental scales. *J. Geophys. Res. Atmos.*, **122**, 5237–5258, <https://doi.org/10.1002/2016JD026013>.
- Lim, K.-S. S., S.-Y. Hong, J.-H. Yoon, and J. Han, 2014: Simulation of the summer monsoon rainfall over East Asia using the NCEP GFS cumulus parameterization at different horizontal resolutions. *Wea. Forecasting*, **29**, 1143–1154, <https://doi.org/10.1175/WAF-D-13-00143.1>.
- Liu, C., K. Ikeda, G. Thompson, R. Rasmussen, and J. Dudhia, 2011: High-resolution simulations of wintertime precipitation

- in the Colorado headwaters region: Sensitivity to physics parameterizations. *Mon. Wea. Rev.*, **139**, 3533–3553, <https://doi.org/10.1175/MWR-D-11-00009.1>.
- Ma, N., G. Y. Niu, Y. Xia, X. Cai, Y. Zhang, Y. Ma, and Y. Fang, 2017: A systematic evaluation of Noah-MP in simulating land-atmosphere energy, water, and carbon exchanges over the continental United States. *J. Geophys. Res. Atmos.*, **122**, 12 245–12 268, <https://doi.org/10.1002/2017JD027597>.
- Maggioni, V., P. C. Meyers, and M. D. Robinson, 2016: A review of merged high-resolution satellite precipitation product accuracy during the Tropical Rainfall Measuring Mission (TRMM) era. *J. Hydrometeorol.*, **17**, 1101–1117, <https://doi.org/10.1175/JHM-D-15-0190.1>.
- Mass, C. F., 2008: *The Weather of the Pacific Northwest*. University of Washington Press, 280 pp.
- , D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? The results of two years of real-time numerical weather prediction over the Pacific Northwest. *Bull. Amer. Meteor. Soc.*, **83**, 407–430, [https://doi.org/10.1175/1520-0477\(2002\)083<0407:DIHRPM>2.3.CO;2](https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2).
- Meynadier, R., G. De Coëtlogon, S. Bastin, L. Eymard, and S. Janicot, 2015: Sensitivity testing of WRF parameterizations on air-sea interaction and its impact on water cycle in the Gulf of Guinea. *Quart. J. Roy. Meteor. Soc.*, **141**, 1804–1820, <https://doi.org/10.1002/qj.2483>.
- Michaelides, S. C., 2008: *Precipitation: Advances in Measurement, Estimation, and Prediction*. Springer, 540 pp.
- Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, <https://doi.org/10.1029/97JD00237>.
- Mooney, P. A., C. Broderick, C. L. Bruyère, F. J. Mulligan, and A. F. Prein, 2017: Clustering of observed diurnal cycles of precipitation over the United States for evaluation of a WRF multiphysics regional climate ensemble. *J. Climate*, **30**, 9267–9286, <https://doi.org/10.1175/JCLI-D-16-0851.1>.
- Morrison, H., G. Thompson, and V. Tatarskii, 2009: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Mon. Wea. Rev.*, **137**, 991–1007, <https://doi.org/10.1175/2008MWR2556.1>.
- Ngailo, T. J., N. Shaban, J. Reuder, M. D. S. Mesquita, E. Rutalebwa, I. Mugume, and C. Sangalungembe, 2018: Assessing Weather Research and Forecasting (WRF) model parameterization schemes skill to simulate extreme rainfall events over Dar es Salaam on 21 December 2011. *J. Geosci. Environ. Prot.*, **6**, 36–54, <https://doi.org/10.4236/gep.2018.61003>.
- Niu, G. Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, <https://doi.org/10.1029/2010JD015139>.
- Odon, P., G. West, and R. Stull, 2017: Vancouver winter 2016/17, how bad was it? *CMOS Bulletin SCMO*, Vol. 45, No. 4, Canadian Meteorological and Oceanographic Society, Ottawa, ON, Canada, 9–12, <https://bulletin.cmos.ca/vancouver-fall-winter-201617-arctic-outbreaks/>.
- , —, and —, 2019: Evaluation of reanalyses over British Columbia. Part II: Daily and extreme precipitation. *J. Appl. Meteor. Climatol.*, **58**, 291–315, <https://doi.org/10.1175/JAMC-D-18-0188.1>.
- Orr, A., C. Listowski, M. Couttet, E. Collier, W. Immerzeel, P. Deb, and D. Bannister, 2017: Sensitivity of simulated summer monsoonal precipitation in Langtang Valley, Himalaya, to cloud microphysics schemes in WRF. *J. Geophys. Res. Atmos.*, **122**, 6298–6318, <https://doi.org/10.1002/2016JD025801>.
- Pei, L., N. Moore, S. Zhong, L. Luo, D. W. Hyndman, W. E. Heilman, and Z. Gao, 2014: WRF Model sensitivity to land surface model and cumulus parameterization under short-term climate extremes over the southern Great Plains of the United States. *J. Climate*, **27**, 7703–7724, <https://doi.org/10.1175/JCLI-D-14-00015.1>.
- Pennelly, C., G. Reuter, and T. Flesch, 2014: Verification of the WRF model for simulating heavy precipitation in Alberta. *Atmos. Res.*, **135–136**, 172–192, <https://doi.org/10.1016/j.atmosres.2013.09.004>.
- Pérez, J. C., J. P. Díaz, A. González, J. Expósito, F. Rivera-López, and D. Taima, 2014: Evaluation of WRF parameterizations for dynamical downscaling in the Canary Islands. *J. Climate*, **27**, 5611–5631, <https://doi.org/10.1175/JCLI-D-13-00458.1>.
- Pleim, J. E., 2007: A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: Model description and testing. *J. Appl. Meteor. Climatol.*, **46**, 1383–1395, <https://doi.org/10.1175/JAM2539.1>.
- Pohl, B., J. Crétat, and P. Camberlin, 2011: Testing WRF capability in simulating the atmospheric water cycle over equatorial East Africa. *Climate Dyn.*, **37**, 1357–1379, <https://doi.org/10.1007/s00382-011-1024-2>.
- Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, **53**, 323–361, <https://doi.org/10.1002/2014RG000475>.
- Pu, Z., C. Lin, X. Dong, and S. K. Krueger, 2019: Sensitivity of numerical simulations of a mesoscale convective system to ice hydrometeors in bulk microphysical parameterization. *Pure Appl. Geophys.*, **176**, 2097–2120, <https://doi.org/10.1007/s00024-018-1787-z>.
- Rajeevan, M., A. Kesarkar, S. B. Thampi, T. N. Rao, B. Radhakrishna, and M. Rajasekhar, 2010: Sensitivity of WRF cloud microphysics to simulations of a severe thunderstorm event over Southeast India. *Ann. Geophys.*, **28**, 603–619, <https://doi.org/10.5194/angeo-28-603-2010>.
- Ralph, F. M., E. Sukovich, D. Reynolds, M. Dettinger, S. Weagle, W. Clark, and P. J. Neiman, 2010: Assessment of extreme quantitative precipitation forecasts and development of regional extreme event thresholds using data from HMT-2006 and COOP observers. *J. Hydrometeorol.*, **11**, 1286–1304, <https://doi.org/10.1175/2010JHM1232.1>.
- Rasmussen, R., and Coauthors, 2012: How well are we measuring snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed. *Bull. Amer. Meteor. Soc.*, **93**, 811–829, <https://doi.org/10.1175/BAMS-D-11-00052.1>.
- Rauber, R. M., and Coauthors, 2019: Wintertime orographic cloud seeding—A review. *J. Appl. Meteor. Climatol.*, **58**, 2117–2140, <https://doi.org/10.1175/JAMC-D-18-0341.1>.
- Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, <https://doi.org/10.1175/2007MWR2123.1>.
- , S. J. Cole, R. M. Forbes, R. J. Moore, and D. Boswell, 2009: Use of high-resolution NWP rainfall and river flow forecasts for advance warning of the Carlisle flood, north-west England. *Meteor. Appl.*, **16**, 23–34, <https://doi.org/10.1002/met.94>.
- Rossa, A., P. Nurmi, and E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts.

- Precipitation: Advances in Measurement, Estimation and Prediction*, Springer, 419–452.
- Sikder, S., and F. Hossain, 2016: Assessment of the weather research and forecasting model generalized parameterization schemes for advancement of precipitation forecasting in monsoon-driven river basins. *J. Adv. Model. Earth Syst.*, **8**, 1210–1228, <https://doi.org/10.1002/2016MS000678>.
- Siuta, D., G. West, and R. Stull, 2017: WRF hub-height wind forecast sensitivity to PBL scheme, grid length, and initial condition choice in complex terrain. *Wea. Forecasting*, **32**, 493–509, <https://doi.org/10.1175/WAF-D-16-0120.1>.
- Skamarock, W., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-468+STR, 113 pp., <https://doi.org/10.5065/D6DZ069T>.
- Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle land surface model (RUC LSM) available in the weather research and forecasting (WRF) model. *Mon. Wea. Rev.*, **144**, 1851–1865, <https://doi.org/10.1175/MWR-D-15-0198.1>.
- Stergiou, I., E. Tagaris, and R.-E. P. Sotiropoulou, 2017: Sensitivity assessment of WRF parameterizations over Europe. *Proceedings*, **1**, 119, <https://doi.org/10.3390/ecas2017-04138>.
- Stull, R., and Coauthors, 2004: Multimodel fine-resolution ensembles for short-range forecasts in mountainous terrain. *20th Conf. on Weather Analysis and Forecasting/16th Con. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., J6.2, https://ams.confex.com/ams/84Annual/techprogram/paper_68296.htm.
- Sun, Q., C. Miao, Q. Duan, H. Ashouri, S. Sorooshian, and K. L. Hsu, 2018: A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Rev. Geophys.*, **56**, 79–107, <https://doi.org/10.1002/2017RG000574>.
- Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, <https://doi.org/10.1175/2008MWR2387.1>.
- Toride, K., Y. Iseri, A. M. Duren, J. F. England, and M. L. Kavvas, 2019: Evaluation of physical parameterizations for atmospheric river induced precipitation and application to long-term reconstruction based on three reanalysis datasets in Western Oregon. *Sci. Total Environ.*, **658**, 570–581, <https://doi.org/10.1016/j.scitotenv.2018.12.214>.
- Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527–548, [https://doi.org/10.1175/1520-0493\(1997\)125<0527:TRDOEM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2).
- Wiersema, D. J., K. A. Lundquist, and F. K. Chow, 2020: Mesoscale to microscale simulations over complex terrain with the immersed boundary method in the Weather Research and Forecasting model. *Mon. Wea. Rev.*, **148**, 577–596, <https://doi.org/10.1175/MWR-D-19-0071.1>.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 676 pp.
- Willmott, C. J., 1981: On the validation of models. *Phys. Geogr.*, **2**, 184–194, <https://doi.org/10.1080/02723646.1981.10642213>.
- Wong, M., G. Romine, and C. Snyder, 2020: Model improvement via systematic investigation of physics tendencies. *Mon. Wea. Rev.*, **148**, 671–688, <https://doi.org/10.1175/MWR-D-19-0255.1>.
- Yang, Z. L., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *J. Geophys. Res.*, **116**, D12110, <https://doi.org/10.1029/2010JD015140>.
- Zeyaeayan, S., E. Fattahi, A. Ranjbar, M. Azadi, and M. Vazifiedoust, 2017: Evaluating the effect of physics schemes in WRF simulations of summer rainfall in North West Iran. *Climate*, **5**, 48, <https://doi.org/10.3390/cli5030048>.
- Zhang, F., M. Li, A. C. Ross, S. B. Lee, and D. L. Zhang, 2017: Sensitivity analysis of Hurricane Arthur (2014) storm surge forecasts to WRF physics parameterizations and model configurations. *Wea. Forecasting*, **32**, 1745–1764, <https://doi.org/10.1175/WAF-D-16-0218.1>.
- Zheng, Y., K. Alapaty, J. A. Herwehe, A. D. Del Genio, and D. Niyogi, 2016: Improving high-resolution weather forecasts using the Weather Research and Forecasting (WRF) Model with an updated Kain–Fritsch scheme. *Mon. Wea. Rev.*, **144**, 833–860, <https://doi.org/10.1175/MWR-D-15-0005.1>.